Ronde 7

Bart Deygers & Marieke Vanbuel
Universiteit Gent
Contact: Bart.Deygers@ugent.be
Marieke.Vanbuel@ugent.be

Hoe ontwikkel je een eerlijke toets voor diverse NT2-profielen?

1. Inleiding

In Nederland bestaan er al decennia gestandaardiseerde toetsen Nederlands als Tweede Taal (NT2) voor migranten, in het bijzonder in het kader van integratieverplichting. Sinds kort hebben ze ook in het Vlaamse onderwijs hun intrede gedaan. Een probleem dat alle gestandaardiseerde toetsen gemeen hebben, is het feit dat ze – wel – standaard zijn. In alle gestandaardiseerde toetsen – ook in zogenaamde '*computer-adaptive* tests' – krijgen kandidaten zoveel mogelijk dezelfde vragen voorgeschoteld om zo weinig mogelijk meetfouten mee te nemen in de meting. Maar laat NT2-leerders nu een bijzonder diverse groep zijn. Ze komen niet alleen uit heel uiteenlopende landen met verschillende moedertalen, maar verschillen ook aanzienlijk in leeftijd, gezinssamenstelling, geletterdheid en scholingsgraad (zie o.a. De Niel e.a. 2016). Hoewel onderzoek doorgaans weinig aandacht besteedde aan die laatste twee factoren, laten recente studies steeds vaker zien dat ze een grote impact hebben op toetsprestaties (Altherr Flores 2021; Deygers e.a. 2022; Deygers & Vanbuel 2022; Helland Gujord 2022). Dat roept allerlei vragen op. Wat gebeurt er als je een toets maakt zonder rekening te houden met een belangrijke maar relatief onzichtbare groep leerders? Kan je toets dan nog construct-relevant zijn? Kan ze nog eerlijk zijn?

In deze tekst bekijken we de impact van geletterdheidsgraad en scholingsniveau op NT2-toetsprestaties. Onze focus ligt daarbij niet zozeer op de prestatie van de leerders op die toetsen, maar op de kwaliteit van de toetsen voor de leerders. In wat volgt, bekijken we eerst de concepten 'construct-relevantie' en 'bias' (twee centrale voorwaarden voor goede taaltoetsen), waarna we het hebben over hoe geletterdheid en scholingsgraad tot meetfouten zouden kunnen leiden, en hoe je valkuilen kan vermijden.

2. Eerlijke en construct-relevante NT2-toetsen

Een toets meet altijd iets. Het construct is datgene wat je wilt meten, graag concreet gemaakt in een definitie. Door een construct expliciet te definiëren, wordt het makkelijker om later te beslissen welke vragen je kunt stellen en welke vragen buiten het

toetsconstruct vallen. Een construct helpt je met andere woorden om te bepalen wat construct-relevant is en wat niet. Daarom maak je je construct best concreet. Een voorbeeld: "Leesvaardigheid is de vaardigheid om tekstuele informatie in digitale vorm doelgericht te begrijpen, af te leiden en te evalueren binnen een brede socio-culturele context".

Relevante vragen bij een construct maken een toets nog niet construct-relevant. Meetfouten treden altijd op, en ze zorgen voor 'construct-irrelevante variatie'. Er zijn twee soorten meetfouten: systematische en niet-systematische meetfouten. De niet-systematische meetfouten zijn niet toe te schrijven aan een designfout van de toets, maar zijn het gevolg van random factoren die er soms op individueel niveau voor zorgen dat iemands score geen goed beeld geeft van iemands vaardigheid. Een kandidaat kan onderpresteren door slechte nachtrust of ziekte, of kan nadeel ondervinden van een bepaald onderwerp in een leestest. Niet-systematische meetfouten zijn willekeurig en je kan ze moeilijk corrigeren in toetsanalyse. Systematische meetfouten daarentegen zijn weeffouten in het toetsdesign en komen terug met een zekere graad van voorspelbaarheid. Systematische meetfouten kunnen verschillende oorzaken hebben, waaronder slecht geconstrueerde vragen (bijvoorbeeld: een meerkeuzevraag met slecht ontwikkelde afleiders), construct-irrelevante vragen (bijvoorbeeld: een pure woordenschatvraag in een leestoets) en oneerlijke vragen.

Oneerlijke vragen zijn vragen die een systematisch en voorspelbaar construct-irrelevant voordeel of nadeel opleveren voor specifieke groepen. Vragen die oneerlijk zijn volgens deze definitie, geven blijk van 'bias'; ze zorgen voor een oneigenlijk construct-irrelevant voordeel of nadeel voor bepaalde groepen. Natuurlijk zijn niet alle systematische scoreverschillen een teken van construct-irrelevantie. Het is bijvoorbeeld perfect mogelijk dat sprekers van een bepaalde moedertaal of met een bepaalde opleidingsgraad consequent hogere scores halen op een bepaalde toets omwille van construct-relevante redenen. Maar wanneer een bepaalde demografische groep consequent nadeel ondervindt van een bepaalde manier van toetsen, dan spreek je van oneerlijkheid.

3. De behoeften van kortgeschoolde, laaggeletterde NT2-leerders

Een heel belangrijke variabele in de brede populatie van NT2-leerders is de combinatie scholingsgraad-geletterdheid. Hoewel het natuurlijk twee verschillende concepten zijn, bekijken we ze in dit hoofdstuk samen, omdat laaggeletterdheid vaak (maar niet altijd) samengaat met minder scholing, en omdat de effecten op toetsprestatie niet altijd makkelijk toe te schrijven zijn aan of geletterdheid of scholingsgraad.

De impact van scholingsgraad-geletterdheid op een toetsprestatie kan groot zijn, maar ze is niet noodzakelijk construct-relevant. De mate van scholing en alfabetische geletterdheid heeft immers een directe impact op niet-talige vaardigheden die essentieel zijn

om toetsen tot een goed einde te brengen. Zo heeft een hogere graad van geletterdheid en scholing een positieve impact op onder andere:

- de verwerking van informatie en instructies;
- het kortetermijngeheugen;
- omgang met hypothetische en abstracte informatie;
- interpretatie van illustraties;
- de verwerking van fonemen;
- teststrategieën inzetten;
- probleemoplossend vermogen.

Deze opsomming maakt duidelijk dat het niet mogelijk is om te veronderstellen dat er geen systematische meetfout zal optreden wanneer een toets die ontwikkeld is voor een hogergeschoold publiek (een diploma secundair onderwijs of hoger) ingezet wordt bij mensen met een lagere graad van scholing en/of geletterdheid (maximaal een diploma lager onderwijs). Als er inderdaad sprake is van bias ten nadele van deze mensen, dan zal de toetsscore een onderschatting zijn van de taalvaardigheid van mensen met dit kwetsbare profiel.

4. Hoe maak je een toets eerlijk voor diverse profielen?

Een inclusieve toets maken die rekening houdt met de noden van mensen met een lagere scholingsgraad-geletterdheid, is natuurlijk enkel nodig wanneer die mensen tot de toetspopulatie behoren (bijvoorbeeld: bij inburgeringstoetsen, brede NT2-instaptoetsen of wetenschappelijke meetinstrumenten voor een breed publiek). Anderzijds willen we wel benadrukken dat deze eerder kwetsbare groep leerders niet klein is. Naar schatting één op de drie NT2-leerders in Vlaanderen heeft een dergelijk profiel.

In een kwalitatief onderzoek (Vanbuel & Deygers te verschijnen) bevroegen we 47 NT2-leerders met diverse graden van scholing-geletterdheid over hoe ze omgaan met bepaalde vragen en vraagtypes op een gestandaardiseerde digitale taaltest. Op basis van de interviews met die deelnemers en op basis van de bestaande literatuur komen we tot de volgende tips voor het ontwikkelen van taaltoetsen die het risico op systematische meetfouten op basis van scholingsgraad-geletterdheid minimaliseren.

- *Houd het aantal vraagtypes beperkt*
  Te veel variatie in vraagtypes kan leiden tot onderpresteren bij kortgeschoolde leerders, omdat zij moeite kunnen hebben met verschillende formats.
  Uit ons onderzoek bleek dat NT2-leerders met een kwetsbaar profiel moeite kunnen hebben met het omschakelen van het ene vraagtype naar het andere.

- *Zorg voor toegankelijke prompts*
  Vragen en inputmateriaal die altijd zichtbaar zijn voor kandidaten, helpen leerders met een lager werkgeheugen om de instructies beter te volgen en de taak correct uit te voeren. Zorg er ook voor dat luisterfragmenten herhaaldelijk kunnen worden afgespeeld.
  Uit de analyse van de interviews bleek dat NT2-leerders met een lagere scholingsgraad of geletterdheid baat kunnen hebben bij herbeluisteren en dat ze dat ook vaker doen dan hogeropgeleide leerders. Het nadeel hierbij is dat leerders met minder goed ontwikkelde teststrategieën de neiging hebben om luisterfragmenten helemaal te herbeluisteren, eerder dan zich te focussen op specifieke fragmenten.

- *Meet vaardigheden-onafhankelijk*
  Mogelijk zorgen geïntegreerde vaardigheden (bijvoorbeeld: lees een tekst en schrijf er een samenvatting over) voor moeilijkheden bij mensen met een lagere scholingsgraad-geletterdheid. Wanneer je één specifieke vaardigheid wilt meten, zorg er dan voor dat de instructies of vragen geen beroep doen op andere vaardigheden. Geef kandidaten van een luistertoets dan ook de optie om instructies, vragen en antwoordopties te beluisteren.
  Uit de interviews bleek dat NT2-leerders met een lagere geletterdheidsgraad baat hebben van gesproken instructies bij luistertaken. Enkele kandidaten hadden zelfs een voorkeur voor die gesproken instructies.

- *Gebruik visuele ondersteuning*
  Visuele ondersteuning kan een taak toegankelijker maken doordat informatie op meerdere manieren wordt gepresenteerd. Zorg er dan wel voor dat de afbeeldingen duidelijk en relevant zijn (bijvoorbeeld: foto's, eerder dan iconen). Duidelijk interpreteerbare afbeeldingen geven immers een idee van de context van een tekst of taak. Je moet dan wel even opletten dat de afbeelding het antwoord op een vraag niet weggeeft, want dan kan er construct-irrelevantie optreden.
  Ook in dit geval zagen we sterke verschillen tussen hoe hoger- en lageropgeleide NT2-leerders omgaan met afbeeldingen in een receptieve taaltoets. Voor hogeropgeleide leerders waren afbeeldingen aangenaam, terwijl ze voor leerders met een lagere scholingsgraad echt belangrijk waren voor tekstbegrip.

- *Vermijd hypothetische situaties*
  Gebruik geen hypothetische situaties in de vragen of prompts (bijvoorbeeld: stel, je schrijft een e-mail aan een vriend), aangezien dit lagergeschoolde leerders kan benadelen. Focus op duidelijke, concrete situaties die makkelijk te begrijpen zijn.

- *Vermijd construct-irrelevante vaardigheden of kennis*
  Sommige vraagtypes doen meer beroep op cognitieve vaardigheden of op kennis van de wereld dan andere. Uit het onderzoek bleek bijvoorbeeld dat hogergeschoolde NT2-leerders voordeel haalden uit een vraag die peilde naar wat niet nodig was

op basis van een bepaalde tekst. Die deductieve vaardigheden waren beter ontwikkeld bij een hogeropgeleide groep. Daarnaast bleek dat vragen op basis van complex gestructureerde informatie in een tabel voor extra moeilijkheden bij kwetsbare NT2-leerders konden zorgen. De lay-out van de informatie en de informatiedichtheid creëerden een probleem.

5. Werkt het? Resultaten uit onderzoek

Na de interviews hebben we de toets in ons onderzoek aangepast. Uit de analyse van die aangepaste versie blijkt dat de toets inderdaad eerlijker wordt voor leerders met een lagere scholingsgraad-geletterdheid. De bias in de toets werd sterk gereduceerd, maar een kleinere bias betekent niet dat de prestaties van verschillende groepen nu gelijk zullen zijn. Ze zijn nu enkel meer construct-relevant, maar dan nog is de impact van bepaalde teststrategieën (die hogeropgeleide leerders sowieso hebben) niet uit te sluiten en kunnen er zich verschillen in prestaties voordoen.

Referenties

Altherr Flores, J. (2021). "The interplay of text and image on the meaning-making processes of adult L2 learners with emerging literacy: Implications for test design and evaluation frameworks". In: *Language Assessment Quarterly*, 18 (5), p. 508-529.

Deygers, B. & M. Vanbuel (2022). "Gauging the impact of literacy and educational background on receptive vocabulary test scores". In: *Language Testing*, 39 (2), p. 191-211.

Deygers, B., M. Vanbuel & U. Knoch (2022). "Can L2 course duration compensate for the impact of demographic and educational background variables on second language writing development?". In: *System*, 109. Online raadpleegbaar op: https://www.sciencedirect.com/science/article/abs/pii/S0346251X22001452.

Helland Gujord, A. (2022). "Who succeeds and who fails? Exploring the role of background variables in explaining the outcomes of L2 language tests". In: *Language Testing*, 40 (2), p. 227-248